Validating Protein Structure Models Using Internal Energy

By: Branden Lee and Kim Kwan

ECS 129 Option 5

<https://github.com/UC-Davis-ECS-129-Project/Protein-Internal-Energy>

Brad:

We should write this project as if it was an actual paper we wanted to publish, so we can get writing experience.

Background questions:

1. Why are we answering this? What is the scientific question we are trying to answer?

2. What is currently known about this topic? Who has already worked on this?

3. What have you done?

Abstract

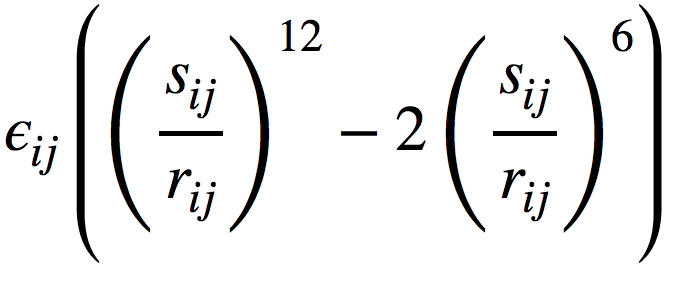
This project seeks to determine which structure of a protein is more likely to be found by comparing internal energy scores. Internal energy scores are calculated by making approximations of the summations that include Van der Waals forces, Coulomb forces, and solvation energy. The project is run in standard python 3 with preprocessed protein data files.

Background

The human body requires proteins to carry out structural, enzymatic, and transport functions in the human body. Currently, we have methods of analyzing structures of proteins such as Edman degradation to analyze the primary structure of a sequence of peptides. We also have protein gels and column chromatography to allow us to analyze certain aspects such as the polarity or size of the protein researchers wish to study. Other methods used to study secondary and tertiary structure include circular dichroism, X-ray crystallography, and NMR spectrometry. However, given a polypeptide sequence, we want to accurately identify the resulting structure based on secondary structures that give off a certain amount of energy based on its position in the overall protein

Intermolecular forces such as Van der Waals forces, Coulomb, and solvation energy have a profound effect on protein folding. Gibbs free energy and spontaneity dictates that a molecule wants to reduce its internal energy, therefore staying in a lower energy conformation. By predicting internal energy of multiple possibilities for 3D protein structures, one can identify the most probable folded structure by finding the structure that minimizes internal energy.

The approximation for Van der Waals energy is done using the equation for Lennard-Jones-Potential as shown in figure 1.



The Lennard-Jones potential is a mathematically simple model that approximates the interaction between a pair of neutral atoms or molecules. A form of this interatomic potential was first proposed in 1924 by John Lennard-Jones. ε is the depth of the potential well, sij is the distance at which the potential reaches its minimum, and r is the distance between the particles. This equation accounts for the attraction and repulsive forces that an atom may experience depending on its distance relative to other atoms within the peptide.

TALK ABOUT THE OTHER PARTS OF THE EQUATION THAT ARE MISSING HERE:

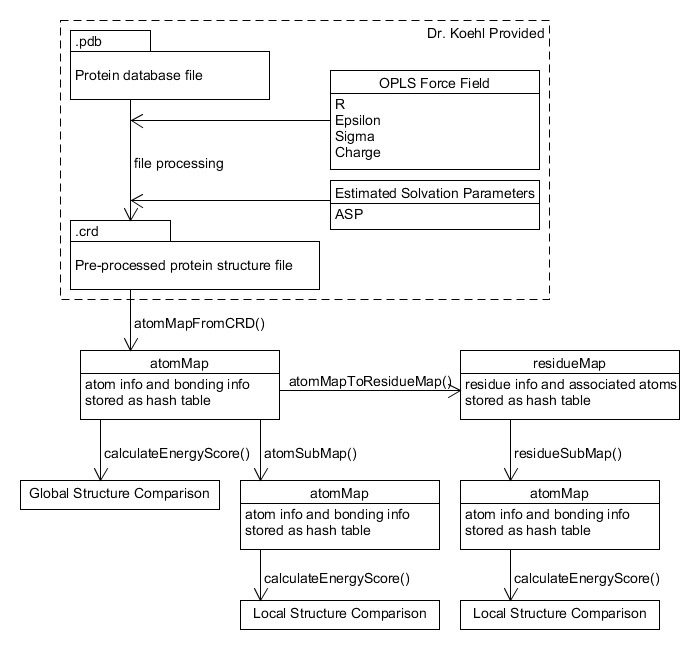
We would also like to determine the presence and location of misalignments and incorrect folding patterns that may have a small or huge impact on the protein in question. While computer models have aided researchers to predict protein models, algorithms built to analyze sequences are not perfect and constantly change over time.

Currently, scientists in this field have been able to detect variations of the same protein to determine the most likely structure that the protein may have with very little variability. However, not much research has been done about finding accurate sequences from two different proteins and observing how they aligned with each other. We are not yet able to observe the variability from the sequence and the model compared to the protein in vivo. We plan to take a sequence that one may have to code for amino acids in a protein and accurately determine which of the two forms given is more energetically favorable.

By predicting accurate models, researchers would be able to identify minute differences between distantly related models. We could use this data to observe how the structure of a protein changes due to mutations in the sequence. Once a library of models is made, professionals in the medical field can access this database and see which specific variations of protein they may need to target. Medical professionals may also use this data to determine if a person is able to react to certain viruses or bacteria that may enter the immune system. They may also use this information to target enzymes that regulate or are a part of many metabolic pathways.

SOLVATION ENERGY NEEDS TO BE DISCUSSED!!

Methods

An internal energy calculator was designed with python. The script opens a preprocessed protein file that contains a tabularized list of atoms in the protein with their associated numerically defined properties. The atoms are stored as a python dictionary and are looped through to calculate internal energy based on the atomic interactions. The program has a time complexity of due to the nested loop.

The pre-processed protein structure file must be in the following format:

Line 1: number of atoms or

Lines with leading pound (#) character will be ignored and not interfere with atom count. Leading pound is used for in-file annotations and comments such as column labels.

The next lines contain rows of atom data with columns delimited by whitespace. Columns are not fixed width. Column width is determined by data type size.

Atom data columns:

|  |  |  |
| --- | --- | --- |
| Column | Data Type | Description |
| 1 | Integer | Atom number |
| 2 | Real(10.4) | X |
| 3 | Real(10.4) | Y |
| 4 | Real(10.4) | Z |
| 5 | Real(10.4) | R |
| 6 | Real(10.4) | Epsilon |
| 7 | Real(10.4) | Sigma |
| 8 | Real(10.4) | Charge |
| 9 | Real(10.4) | ASP |
| 10 | Char(6) | Atom name |
| 11 | Char(6) | Residue name |
| 12 | Integer | Residue number |

The next lines contain rows of atom bonding data with columns delimited by whitespace.

Atom bonding data columns:

|  |  |  |
| --- | --- | --- |
| Column | Data Type | Description |
| 1 | Integer | Atom number |
| 2 | Integer | Size of subsequent integer array |
| 3 | Integer Array | Bonded atom number |

Output: Energy levels of each protein model after calculations. Also, printing of statement that tells the user which protein model works best depending on energy levels

Implementation/Code:

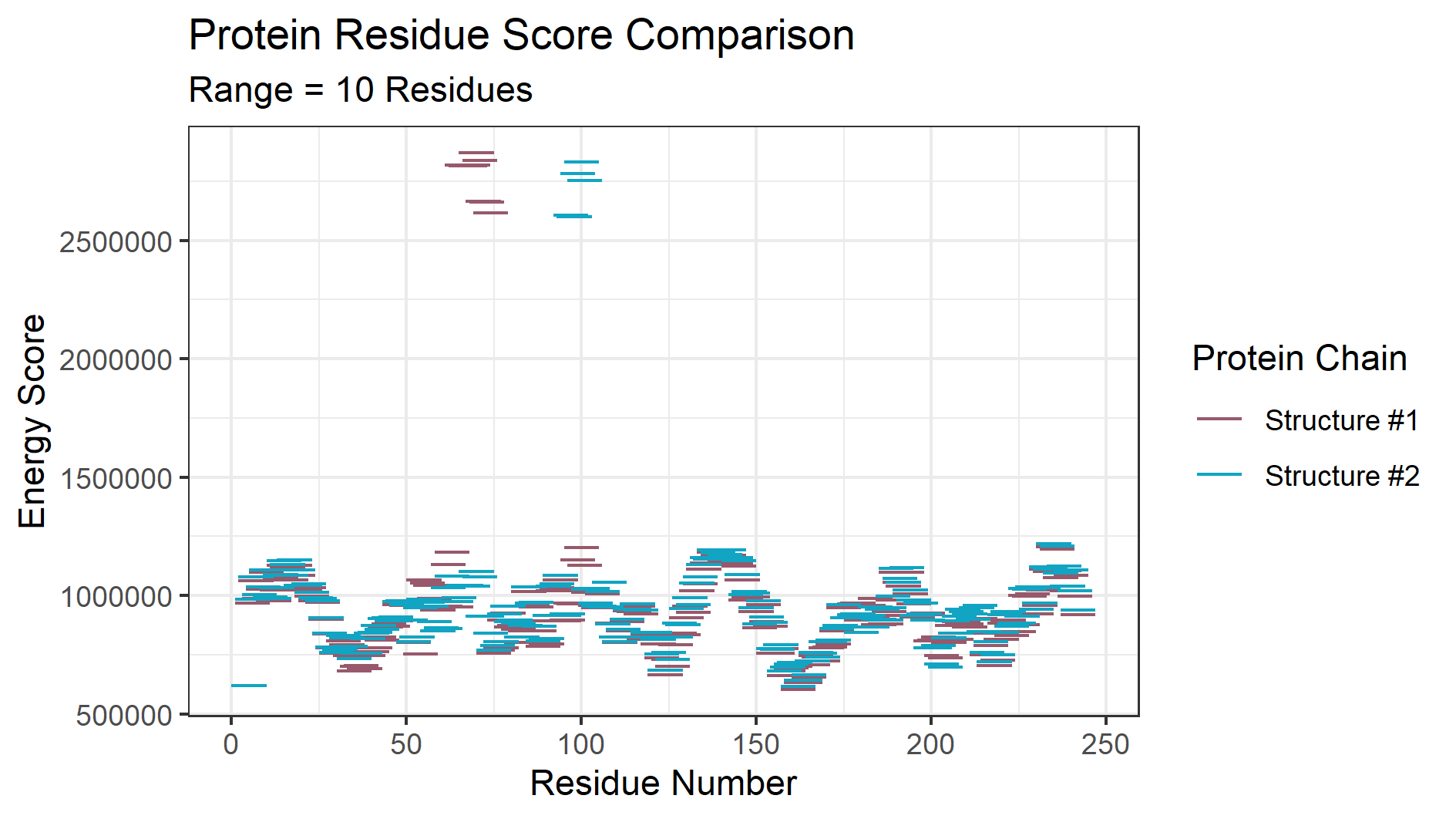
1. Separate the code and store them into lists
2. Calculate energy level based on equation given to use in figure 1
3. ???somehow analyze data to determine energy levels????
4. Compare energy levels
5. Tell user which protein structure is more likely based on energy level

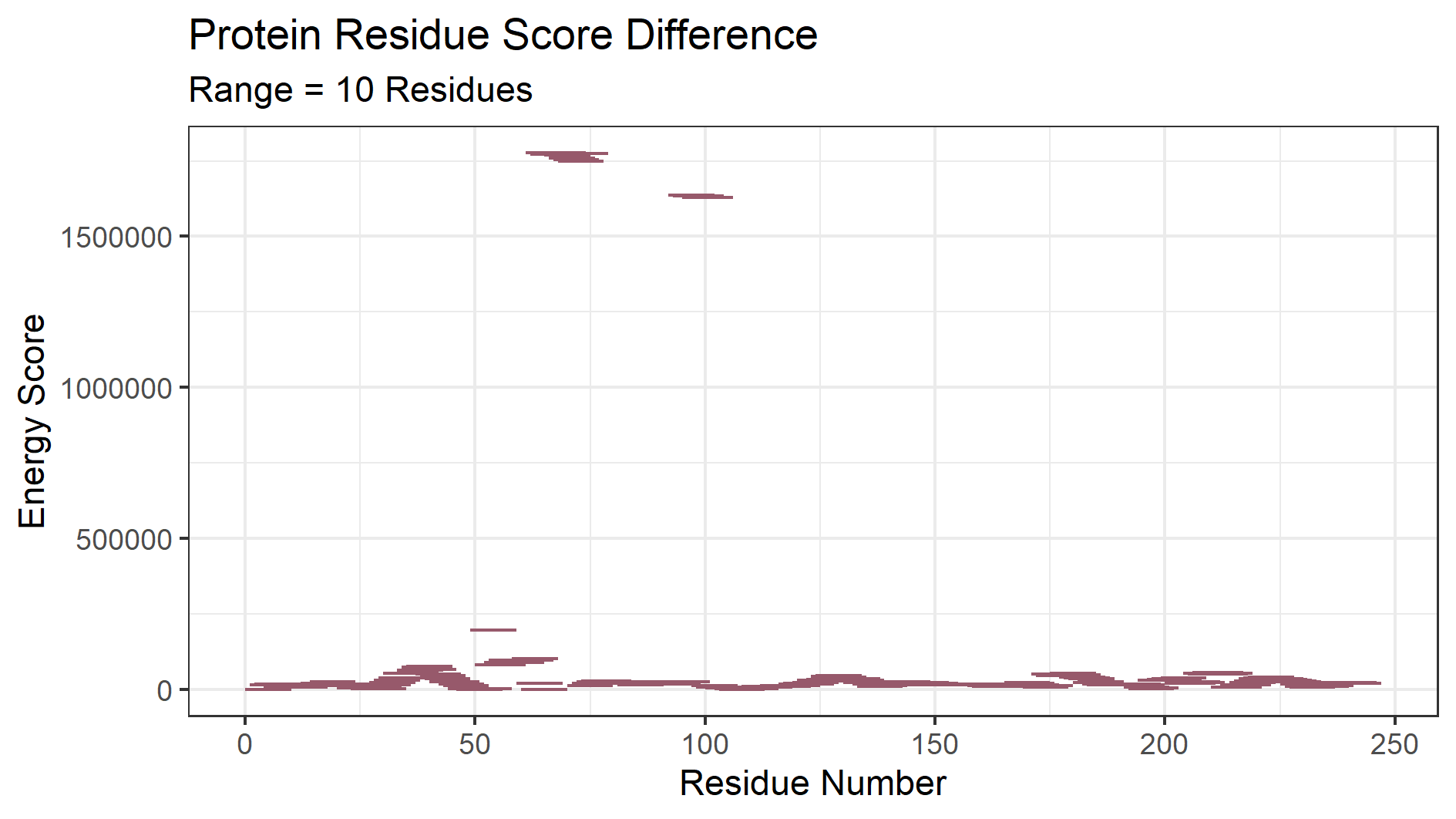
(Page Break)

Results

The energy score of structure #1 was , while conformation #2 was. There was a significant difference in the energy scores between both protein conformations. structure #2 energy score is lower and is the more likely conformation because of thermodynamic reasons.

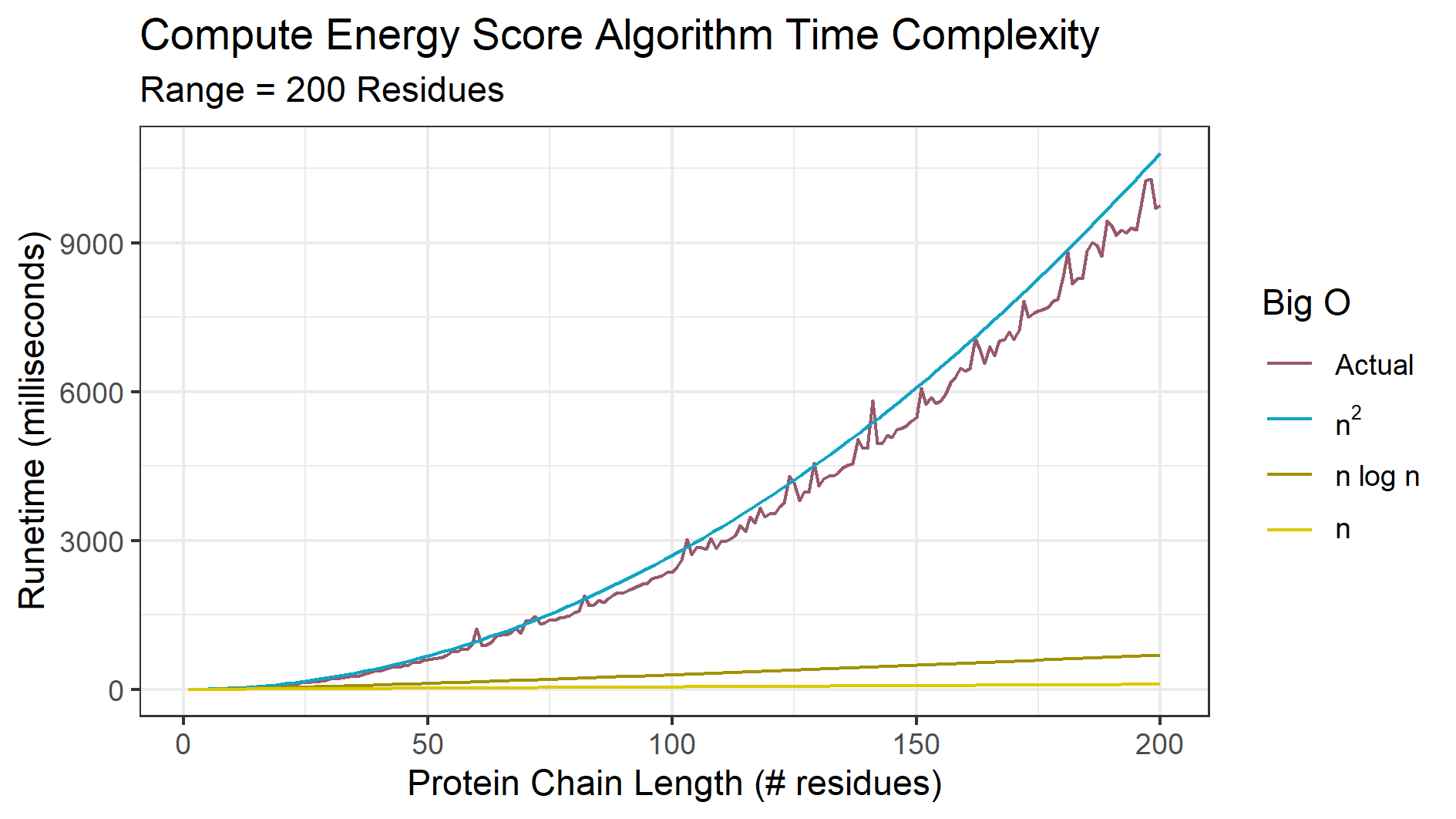
Both conformations of the protein had an average energy score difference ~10 Mcal/mol; however, from amino acid 66 to 69, there was a significant difference. A subarray range of five residues was used to discover the local energy scores.

Fig 4.1

 Fig 4.2

(Image of Protein Structure #1 and #2 superimposed)

The Lennard-Jones potential and electrostatic energy calculations are in a nested loop, thus the time complexity of the algorithm is theorized as . Running the protein energy scoring algorithm on randomly generated protein chains of length up to confirms that the algorithm runs at a speed.



Discussion

The program takes around 7 seconds which is quite lengthy. With a sufficiently large amount of possible protein topologies for a single protein that need to be validated, the script could take a large amount of time to finish. A compiled language such as C would be a better fit.

Works Cited

Breda A, Valadares NF, Norberto de Souza O, et al. Protein Structure, Modelling and Applications. 2006 May 1 [Updated 2007 Sep 14]. In: Gruber A, Durham AM, Huynh C, et al., editors. Bioinformatics in Tropical Disease Research: A Practical and Case-Study Approach [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2008. Chapter A06. Available from: https://www.ncbi.nlm.nih.gov/books/NBK6824/

Maiorov and Abagyan, 1998 V. Maiorov, R. Abagyan Energy strain in three-dimensional protein structures Fold. Des., 3 (1998), pp. 259-269 <https://www.sciencedirect.com/science/article/pii/S1359027898000376>

UCSF Chimera--a visualization system for exploratory research and analysis. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. J Comput Chem. 2004 Oct;25(13):1605-12.

Chang, Raymond. *Physical Chemistry for the Biosciences.*Sausalito, CA. University Science Books, 2005. (498-500)